Optimally Efficient Estimation of the Statistics of Rare Events in Queueing Networks

Michael R. Frater, Tava M. Lennon, and Brian D. O. Anderson, *Fellow, IEEE*

Abstract—**Because of their rarity, the estimation of the statistics of buffer overflows in networks of queues by direct simulation is very costly. An asymptotically optimal (as the overflow recurrence time becomes large) scheme has been proposed by others, using importance sampling. This paper addresses two aspects of this scheme. First, in the existing approach, a numerical minimization is required to generate the simulation network. This paper describes an equivalent analytic minimization. A simple procedure for constructing the optimal simulation network is included. Second, it is shown that the average behavior of the simulation system is the same as the average behavior of the original network in the period leading up to a buffer overflow.**

I. Introduction

THE efficient estimation of the statistics of rare events has been of interest for a number of years [1]–[4]. The large deviations theory has been applied to perform asymptotically optimal (in the sense of variance) simulations of rare events [3]. More recently, specific application of this theory has been made to buffer overflows in queueing networks [5]–[7], in which the large deviations theory and numerical techniques have been used to find optimally fast simulation systems for networks of queues. In [4], linear algebraic techniques are used to find the distribution of the quantities of interest. While the methods of [4] have the advantage of being more direct, and of yielding distribution functions rather than just expected values, they involve the inversion of a probability transition matrix, and we believe that the fast simulation methods have computational advantages in problems of large dimension. Queueing networks are an example of such a problem, where the order of the transition matrix is N^d, where N is the buffer size and d is the number of queues in the network.

This paper addresses two aspects of the large deviations approach to finding optimal simulation systems for queueing networks. The first relates to the finding of the optimal simulation system, in which a numerical minimization has been required previously. A simple direct analytic solution to this problem is presented here. The solution to this problem has been presented previously for tandem networks of queues [7]. This paper demonstrates that a unique solution to the minimization problem exists for more general queueing networks. The construction of such networks is also discussed. Second, the behavior of the optimal simulation system generated is related to that of the original system leading up to a buffer overflow. This result is analogous to results previously obtained for diffusions [8].

The theory used in this paper is outlined in Section II. The analytic solution of the minimization problem is set out in Section III, (the minimality of this solution is proved in Appendix B.) In Section IV, the relationship between the behavior of the optimal simulation system and that of the original system in the period leading up to an overflow is discussed.

Manuscript received February 23, 1990; revised May 16, 1991. Paper recommended by Associate Editor at Large, P. R. Kumar. This work was supported by the Australian Telecommunications and Electronics Research Board.

M. R. Frater is with the Department of Electrical Engineering, University College, Australian Defence Force Academy, Canberra ACT 2600, Australia.

T. M. Lennon and B. D. O. Anderson are with the Department of Systems Engineering, Australian National University, Canberra ACT 2601, Australia.

IEEE Log Number 9103552.

II. Background Theory

A. Importance Sampling

In a queueing network with finite buffers, a certain proportion of packets are lost due to buffer overflows. While the mean time between overflows can be calculated analytically for a single $M/M/1$ queue (e.g., [6]), the first step equations for a network of queues cannot be solved analytically because the order of the characteristic equation becomes large. Therefore, simulation is often used to find the recurrence time of buffer overflows. Because the mean time between overflows is large in a properly dimensioned network, direct simulation may not be feasible, simply because of its large cost in computer time. However, using the idea of importance sampling, the mean time between overflows can be found by simulation without incurring the large cost involved in direct simulation.

The idea in importance sampling is as follows. Suppose we are interested in certain (rare) events occurring in a system S that we can simulate on a computer. Then instead of simulating S we simulate a second system $\bar{S}$, which has the property that events in S and $\bar{S}$ correspond in some way. In particular, to the rare events A in S correspond events $\bar{A}$ in $\bar{S}$. The correspondence is such that

1) the events $\bar{A}$ in $\bar{S}$ are more frequent than the events A in S, and

2) the connection between S and $\bar{S}$ allows one to infer $P(A)$ if one knows $\bar{P}(\bar{A})$. ($\bar{P}(\bar{A})$ is probability of event $\bar{A}$ in system $\bar{S}$.)

In this paper, the system $\bar{S}$ will be a network of queues.

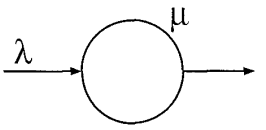


Fig. 1. An $M/M/1$ queue.

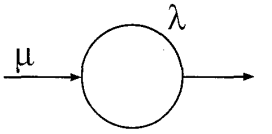


Fig. 2. The optimal simulation system for the $M/M/1$ queue in Fig. 1.

The system $\bar{S}$ will be a network of queues also, with the same structure as S, but with various parameters such as arrival and service rates that will be different from the corresponding quantities in S.

An analytic solution to the problem finding the mean time between overflows from an $M/M/1$ queue has been given previously [6]. Nevertheless, it also serves to illustrate the idea in importance sampling, which can also be used for speeding up the simulation of networks. Take the $M/M/1$ queue shown in Fig. 1 as the system S. The second system $\bar{S}$ we consider is shown in Fig. 2. It is also an $M/M/1$ queue, but with different arrival an service rates. The fact that S and $\bar{S}$ have the same structure with different transition probabilities between states is typical of importance sampling. Notice that in fact the queue in Fig. 2 is unstable. In Fig. 1, A is the event that the queue length, starting at 0, reaches N before returning to 0. In Fig. 2, $\bar{A}$ is the event that the queue length, starting at 0, reaches N before returning to 0. Obviously, the computation of the mean time for $\bar{A}$ to occur simulation will be rapid. From it, one can deduce the mean time for A to occur as we now outline.

It is well known that one can set up an embedded discrete-time Markov chain $\{X_k, k = 0, 1, 2, \cdots, \}$, with X_k denoting the queue length just after the kth change of that length, i.e., just after the kth occurrence of an arrival or departure. For convenience, one can rescale time so that $\lambda + \mu = 1$. Then, λ, μ are, respectively, the probabilities of upward and downward transitions given $X_k > 0$. Where $X_k = 0$, the probability of an upward transition is 1.

Now let α be the probability that with $X_0 = 0$, X_k hits N before hitting zero again. Let T denote the first time X_k reaches N and let J denote the time to hit either 0 or N for the first time after leaving 0. It is not hard to check (see also [6]) that

$$E[T] = \frac{1}{\alpha} E[J] \tag{1}$$

and that α and $E[J]$ are given by

$$\alpha = \frac{1 + \frac{\mu}{\lambda}}{1 - \left(\frac{\mu}{\lambda}\right)^N} \tag{2}$$

$$E[J] = \frac{1 - N\alpha}{\mu - \lambda}. \tag{3}$$

Our interest, however, is to understand how these values can be obtained by (efficient) simulation. Now, $E[J]$ can be easily obtained by direct simulation on S, while α, which will be small, is obtainable from $\bar{S}$, as follows. Let us call a *cycle* of the system S or $\bar{S}$ a movement from 0 to the first time *either* 0 is reached again, *or* N is reached. Define $V_k = 1_{\{X_m \text{ reaches } N \text{ in cycle } k\}}$. For S, we have

$$E[V_k] = \alpha. \tag{4}$$

Let L_k denote the likelihood ratio $dP/d\bar{P}$ during cycle $\boldsymbol{k}$. Notice that the L_k are i.i.d. and

$$\bar{E}[L_k V_k] = E[V_k] = \alpha. \tag{5}$$

Now L_k is readily definable for S and $\bar{S}$. Suppose that $V_k = 1$, and that there are l departures and therefore $N + l$ arrivals in the cycle. (There must be N more arrivals than departures for a cycle starting at 0 to end at N). Let ζ be a trajectory starting at 0, ending at N, with $N + l$ arrivals and l departures in the cycle. Then

$$\begin{aligned} \bar{P}(\zeta) &= \mu^{N+l-1}\lambda^l \\ P(\zeta) &= \lambda^{N+l-1}\mu^l \end{aligned} \tag{6}$$

and so

$$L_k = \frac{dP}{d\bar{P}} = \left(\frac{\lambda}{\mu}\right)^{N-1} \tag{7}$$

on the set $\{V_k = 1\}$.

There are frequent occurences of the set $\{V_k = 1\}$ for the system $\bar{S}$, since it is unstable. We examine p cycles for $\bar{S}$ and estimate α by

$$\hat{\alpha} = \frac{L_1 V_1 + L_2 V_2 + \cdots + L_p V_p}{p} \tag{8}$$

$$= \left(\frac{\lambda}{\mu}\right)^{N-1} \frac{\text{Number of cycles for which } V_k = 1}{p}. \tag{9}$$

The speed up factor in simulation time obtained by using $\bar{S}$ instead of S and requiring equal accuracy in estimating α turns out to be, see [6]

$$\left[N\left(\frac{\lambda}{\mu}\right)^N \left(1 - \frac{\lambda}{\mu}\right)\right]^{-1}. \tag{10}$$

In the above analysis we have shown how knowing and simulating $\bar{S}$, it can be used to compute α and thus important statistics concerning S. But we have not described what led us to the particular $\bar{S}$ chosen, rather than something else. *A major issue in the use of importance sampling is how one should construct $\bar{S}$ from S.* To an extent, the problems of obtaining the probability of a rare event or the mean time between occurrences of a rare event (which are both problems of excessive computer time) are being replaced by another difficult problem (How should we obtain $\bar{S}$ from S?) in importance sampling.

This problem can be posed as an optimization problem, in the following way. Let A be a rare event for S: $\alpha = P(A) \ll 1$. For a direct Monte Carlo simulation involving n independent experiments we could estimate α via

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} 1_A(\omega_i) \tag{11}$$

where the ω_i are the i.i.d. outcomes of the experiments. The variance of $\hat{\alpha}_n$ is easily computed as

$$E\left[\alpha - \hat{\alpha}_n\right]^2 = \frac{1}{n}(\alpha - \alpha^2). \tag{12}$$

Alternatively, consider a probability measure $\bar{P}$ associated with a system $\bar{S}$, with P absolutely continuous with respect to $\bar{P}$. Using $\bar{S}$ we can obtain a second estimate

$$\hat{\bar{\alpha}}_n = \frac{1}{n}\sum_{i=1}^{n} 1_A(\bar{\omega}_i) L(\bar{\omega}_i) \tag{13}$$

where $L = dP/d\bar{P}$ and the $\bar{\omega}_i$ are the outcome of n experiments using $\bar{S}$. The variance of $\hat{\bar{\alpha}}$ is different to (12), and is obtainable as

$$\frac{1}{n}\left(\int_A L^2(\omega)\, d\bar{P}(\omega) - \alpha^2\right). \tag{14}$$

We want this to be as accurate as possible. So we want to adjust all the transition probabilities in S to new ones in $\bar{S}$ so that

$$(\sigma^*)^2 = \int_A L^2(\omega)\, d\bar{P}(\omega) \tag{15}$$

is minimized.

Given a system $\bar{S}$ minimizing $(\sigma^*)^2$, we can use (8) to find the value of α for the original system S from (much faster) simulation performed on $\bar{S}$.

B. Large Deviations Theory

Large deviations theory has been used to obtain a number of asymptotic results [6] that apply not only to Jackson networks, but also to more general queueing networks. Some of these, that are relevant to the results presented below, are summarized here.

Let $\xi_1 \cdots \xi_d$ be i.i.d. random variables in $\mathbb{R}^d$. Let F be the distribution function of the $\{\xi_k\}$ and m its mean. Assume that the Laplace transform of F

$$M(s) = \int_{\mathbb{R}^d} \exp\langle s, z\rangle\, dF(z) \tag{16}$$

is finite in a neighborhood of 0. Then the Cramér or Legendre transform is defined as [3]

$$h(y) = \sup_{s\in\mathbb{R}^d}\left[\langle s, y\rangle - \log M(s)\right]. \tag{17}$$

For example, the Cramér transform of an exponential distribution with parameter λ is

$$h_\lambda(u) = \begin{cases} \lambda u - \log(\lambda u) - 1 & u > 0 \\ \infty & \text{otherwise.} \end{cases} \tag{18}$$

The following properties of the Cramér transform are used in this paper:

1) $h(\cdot)$ is convex;
2) $h(\cdot)$ is nonnegative;
3) $h(y) = 0$, if and only if $y = m$, where m is the mean of the distribution function F;
4) $h'(m) = 0$.

Properties 3 and 4 are easily verified for $h_\lambda(u)$ in (18) on setting $u = 1/\lambda$.

For a network of queues, call a *cycle* a piece of a trajectory starting at the zero state and terminating on the first occasion when either the total number of customers in the network exceeds some value (say N), or the state equals zero again. Call a cycle that terminates with the system in the empty state a cycle of the first type, and one that terminates with the number of customers in the network greater than N a cycle of the second type.[1] Let d be the number of queues in the network, λ_i be the rate of external arrivals at queue i, γ_i be the total arrival rate at queue i, μ_i be the virtual service rate at queue i, p_{ij} be the routing probability from queue i to queue j and p_{i0} be the probability that a customer leaving queue i leaves the network. For current purposes, we will assume that all external arrival processes are poisson, that all the service rates are exponentially distributed, and that all queues are asymptotically stable (i.e., $\gamma_i < \mu_i \forall i$). All these parameters of the system S (i.e., γ_i, λ_i, μ_i and p_{ij}) are assumed constant. These parameters of the system satisfy the *traffic equations*

$$\gamma_i = \sum_{j=1}^{d} p_{ji}\gamma_j + \lambda_i \tag{19}$$

and the routing probabilities satisfy

$$\sum_{j=0}^{d} p_{ij} = 1. \tag{20}$$

Suppose α is the probability that a cycle ends in a buffer overflow, (i.e., that it is of the second type). There is a relation between α and a system $\bar{S}(\lambda_i', \mu_i', \gamma_i', p_{ij}')$, which is obtained from S by varying its parameters, and which is used for estimating α by simulation. This relation is derived by heuristic argument in [6]. The parameters for a system $\bar{S}(\lambda_i', \mu_i', \gamma_i', p_{ij}')$ can be found as the arguments achieving minimization in the following large-deviations approximation for α:

$$\alpha \simeq \exp - N \inf_{\lambda_i', \mu_i', \gamma_i', p_{ij}'} R\left[\sum_{i=1}^{d} \lambda_i' h_{\lambda_i}\left(\frac{1}{\lambda_i'}\right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right) + \sum_{i=1}^{d} \min(\gamma_i', \mu_i') K_i\right] \tag{21}$$

where

$$K_i = \sum_{j=0}^{d} p_{ij}' \log \frac{p_{ij}'}{p_{ij}} \tag{22}$$

$$R = \frac{1}{\sum_i (\gamma_i' - \mu_i') 1_{\gamma_i' > \mu_i'}}. \tag{23}$$

[1] A cycle could also be defined as terminating when the number of customers in any one queue exceeded some predetermined level or is again zero [6].

The infimum is subject to the following constraints:

$$\lambda_i', \mu_i', \gamma_i' \geq 0 \tag{24a}$$

$$0' \leq p_{ij}' \leq 1 \tag{24b}$$

$$\gamma_i' > \mu_i' \qquad \text{for at least one } i \tag{24c}$$

$$\sum_{i=1}^{d} (\lambda_i' + \mu_i') = 1 \tag{24d}$$

$$\sum_{j=0}^{d} p_{ij}' = 1 \tag{24e}$$

$$\gamma_i' = \sum_{j=1}^{d} p_{ji}' \min(\gamma_j', \mu_j') + \lambda_i'. \tag{24f}$$

It has been argued [3], [6] that if the system $\bar{S}$ defined by the parameters γ_i', λ_i', μ_i', and p_{ij}' is used to perform simulation, then this simulation is asymptotically optimal as the mean time between overflows grows large, i.e., as N becomes large. We will perform the minimization using the method of Lagrange multipliers to satisfy the equality constraints. The solution obtained will then be shown to satisfy the inequality constraints.

It should be emphasized that the derivation present in [6] is based on heuristic arguments, and that, to the best of our knowledge, a rigorous justification of this result is not yet known. However, it has been shown rigorously that this approach does indeed produce a simulation system that is asymptotically optimal for some systems, such as the $GI/GI/1$ queue [9]. Further, the results of Section IV tend to verify this claim. In this section, we have not presented the arguments of [6] leading to the cost function (21). A summary of these arguments, for the case of a single $GI/GI/1$ queue, are presented in Appendix A.

III. The Optimal Simulation System

In this section, a heuristic motivation for the results obtained is given, followed by a direct analytic solution to the minimization problem described above for Jackson networks, (i.e., we assume that all the external arrival streams are poisson, and that the service times are exponentially distributed.) After the mathematical details of the solution, some comments on interpretation are made. A proof of the optimality of the solution is contained in Appendix B.

A. Heuristic Motivation

It has been shown previously that the optimal simulation system for an $M/M/1$ queue is another $M/M/1$ queue with arrival and service rates reversed, as shown in Figs. 1 and 2 [6], [7].

Consider the case of two queues in tandem, as shown in Fig. 3. Assume $\lambda < \mu_1$ and $\lambda < \mu_2$. If, for example the service rates satisfy $\mu_1 > \mu_2$, and if the size of the buffers is large, then the overflow statistics ought to be dominated by the second buffer's behavior, and μ_1 should be of little importance, i.e., from the point of view of buffer overflows, the system should behave essentially as a single buffer with arrival rate λ and service rate μ_2. Hence, it is reasonable to

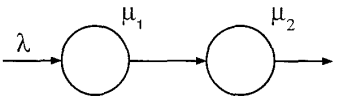


Fig. 3. A pair of queues in tandem.

suspect that for large buffer sizes, the optimal simulation system will behave very like a single buffer with arrival rate μ_2 and service rate λ, even if the difference between μ_1 and μ_2 is small. Similarly, if we have a network of queues, and large buffer sizes, it is reasonable to suspect that the queue with the largest load (i.e., the largest ratio of arrival to service rate) will dominate the overflow statistics, and that the behavior of the simulation system will be similarly affected. Let us now see how this conclusion can be rigorously established, using methods that will apply to arbitrary Jackson networks, and also to some classes of more general queueing networks.

B. Evaluation of $\bar{S}$

In order to find the optimal simulation system, the arguments achieving the infimum in the exponent of (21) must be found, subject to the constraints listed. In order to do this, we define a Lagrangian as follows, with Lagrange multipliers g, b_i, and c_i:

$$\begin{aligned}\mathcal{L} = R\left[\sum_{i=1}^{d} \lambda_i' h_{\lambda_i}\left(\frac{1}{\lambda_i'}\right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right) + \sum_{i=1}^{d} \min(\gamma_i', \mu_i') K_i\right] \\ + g\left(\sum_{i=1}^{d} (\lambda_i' + \mu_i') - 1\right) \\ + \sum_{i=1}^{d} b_i\left(\lambda_i' + \sum_{j=1}^{d} \min(\gamma_j', \mu_j') p_{ji}' - \gamma_i'\right) \\ + \sum_{i=1}^{d} c_i\left(\sum_{j=0}^{d} p_{ij}' - 1\right).\end{aligned} \tag{25}$$

Each of the equality constraints (24d) through (24f) is associated with a Lagrange multiplier. We will assume without real loss of generality that queue 1 has the largest load, (i.e., $\rho_1 > \rho_i$ for all $i \neq 1$, where $\rho_j = \gamma_j/\mu_j$).

Define r_i as the expected number of times that a customer arriving at queue i will pass subsequently through queue 1 before leaving the network.[2] Because the routing is Markovian, r_i does not depend in any way on the previous history of a customer, e.g., whether the customer enters that network at queue i, or comes to queue i from within the network. Thus r_i is also the expected number of visits to queue 1 of a customer entering the network at queue i. Then

$$\sum_{i=1}^{d} r_i \lambda_i = \gamma_1. \tag{26}$$

[2] r_i is easily calculated as the value of γ_1 when the values $\lambda_i = 1$ and $\lambda_j = 0$ for $j \neq i$ are substituted into the traffic equations (19). This is possible because the r_i depend only on the routing in the network, and therefore do not change when the external arrival rates are changed.

We note that $r_i = 0$ implies that customers arriving at queue i can never be routed through queue 1 before leaving the network. Also, since all customers arriving at queue 1 must pass through this queue before leaving the network, we must have $r_1 \geq 1$.

When the derivatives of the Lagrangian with respect to the parameters of the system $\bar{S}$ are evaluated (see Appendix B), it can be shown that the following values of the parameters of the system $\bar{S}$ correspond to a turning point of the Lagrangian, and are in fact the required infimum.

$$\gamma_i' = \gamma_i\left[1 + \frac{r_i}{r_1}\frac{\mu_1 - \gamma_1}{\gamma_1}\right] \qquad \text{(which implies } \gamma_1' = \mu_1\text{)} \tag{27a}$$

$$\gamma_i' = \lambda_i\frac{\gamma_i'}{\gamma_i} \tag{27b}$$

$$\mu_i' = \begin{cases} \gamma_1 + \dfrac{(r_1 - 1)(\mu_1 - \gamma_1)}{r_1} & \text{for } i = 1 \\ \mu_i & \text{for } i > 1 \end{cases} \tag{27c}$$

$$p_{i0}' = p_{i0}\frac{\gamma_i}{\min(\gamma_i', \mu_i')} \tag{27d}$$

$$p_{ij}' = p_{ij}\frac{\gamma_i}{\min(\gamma_i', \mu_i')}\frac{\gamma_j'}{\gamma_j} \qquad \text{for } j > 0. \tag{27e}$$

The Lagrange multipliers take values

$$g = 0 \tag{28a}$$

$$b_i = -R\log\frac{\gamma_i'}{\gamma_i} \tag{28b}$$

$$c_i = R\min(\gamma_i', \mu_i')\left[\log\frac{\min(\gamma_i', \mu_i')}{\gamma_i} - 1\right]. \tag{28c}$$

In the simulation system defined here, we can show that queue 1 becomes unstable in $\bar{S}$, and that all other queues remain stable in $\bar{S}$. From (27a) and (27c), we can see that $\gamma_1' > \mu_1'$ if and only if

$$\mu_1 > \frac{\gamma_1 + (r_1 - 1)\mu_1}{r_1} \tag{29}$$

which is true if and only if $\gamma_1 < 1$, which is given. Hence, queue 1 is unstable in $\bar{S}$.

Next, we show that no other queues are unstable in $\bar{S}$. It has been assumed that $\rho_1 > \rho_j \forall j \neq 1$. Also, we must have $r_1 > r_j$, because all customers arriving at queue 1 are counted in r_1, and it is not possible to have more of these customers counted in r_j. Therefore

$$\rho_1^{-1} - 1 < \rho_j^{-1} - 1. \tag{30}$$

Hence,

$$\frac{\rho_1^{-1}}{r_1} < \frac{\rho_j^{-1} - 1}{r_j} \tag{31}$$

and simple manipulation yields

$$\gamma_j\left[1 + \frac{r_j}{r_1}\frac{\mu_1 - \gamma_1}{\gamma_1}\right] < \mu_j \qquad \forall j \neq 1, \tag{32}$$

i.e., substituting from (27a) and (27c), we must have

$$\gamma_j' < \mu_j' \qquad \text{for } j \neq 1. \tag{33}$$

In other words, all queues are stable in the optimal simulation system except for queue 1.

This instability of queue 1 satisfies the last of the inequality constraints in (24) (that $\gamma_i' > \mu_i'$ for some i). It is shown in the appendix that the other inequality constraints are satisfied.

C. Interpretation

While no physical understanding is necessary to generate an asymptotically optimal simulation system using the above equations, it is nonetheless useful to see the meaning of the above transformation.

- Only the dominating queue (i.e., the queue with the largest load) becomes unstable in the simulation system, as was shown above.
- $r_i = 0$ implies that no customers passing through queue i reach queue 1, and, from (27a), that $\gamma_i' = \gamma_i$. Hence, (27b) implies that $\lambda_i' = \lambda_i$, i.e., arrival rates are changed only at external inputs from which customers may be routed to the dominating queue.
- The arrival rate at the dominating queue in the simulation system is always the same as the service rate of this same queue in the original system.
- The rate of customers leaving the network at external outputs remains unchanged in the simulation system (27d). This can be seen in (27d), where it is clear that $p_{i0}' \min(\gamma_i', \mu_i') = p_{i0}\gamma_i$.
- Only those parts of the network that can contribute to contribute to overflows in the original system contribute to overflows in the simulation system. That is, it is possible for customers to be routed from one queue to another in the simulation system if and only if it is possible in the original system. This can be seen from (27e), where it is clear that $p_{ij}' > 0$ if and only if $p_{ij} > 0$.
- We note that the solution obtained conforms to the heuristic ideas set out in Section III-A.

It should also be noted (perhaps surprisingly) that the distributions of service times for queues other than that dominating the overflow statistics are not required to be exponential, and the external arrival processes at queues from which there is not direct path to the input of queue 1 (i.e., $r_j = 0$) need not be poisson. Both of these facts are demonstrated in Appendix B.

IV. Behavior of Systems Leading to Overflow

In this section, it will be shown that there is a connection between the average behavior of the original system S in the time leading up to an overflow, and the average behavior of the simulation system $\bar{S}$ described above. In order to do this, the reverse-time model of the network will be used. This is defined in Section IV-A, and is followed by a comparison of

the behavior of S and $\bar{S}$ for an isolated $M/M/1$ queue (4.2) and for a Jackson network (4.3).

A. Calculation of Reverse-Time Model

Given a stationary finite-state Markov process defined by the transition probabilities

$$q_{ij} = P\big(x(t+1) = j \mid x(t) = i\big) \tag{34}$$

its reverse-time model is defined by the transition probabilities ([10, p. 28])

$$\tilde{q}_{ji} = P\big(x(t) = i \mid x(t+1) = j\big) \tag{35}$$

where

$$\tilde{q}_{ji} = \frac{\pi(i)}{\pi(j)} q_{ij} \tag{36}$$

and $\pi(\cdot)$ is the invariant probability of a state. (Some readers may simply recognize (36) as Bayes rule, given stationarity.) The natural direction of time flow for the reverse-time model is backwards (i.e., with the time index t decrementing), under which it has the same invariant probability as the forwards-time system (whose time index t increases.)

The reverse system traces out the same trajectories with the same probability as the forward system, but does so backwards in time. Hence, the expected time to travel from A to B in the forward system is the same as the expected time to travel from B to A in the reverse system.

A system is said to be reversible if $\tilde{q}_{ji} = q_{ji}$ [10]. An $M/M/1$ queue is an example of a reversible system.

We now turn our attention specifically to Jackson networks. Let $(\gamma_{iR}, \mu_{iR}, \lambda_{iR}, p_{ijR})$ be the parameters of the reverse-time network corresponding to $(\gamma_i, \mu_i, \lambda_i, p_{ij})$ in the forwards-time network. We assume that the forward system is asymptotically stable, (i.e., the arrival rate γ_i is less than the service rate μ_i at all queues.)

The reverse-time model of a queueing network with infinite buffers is easily found. In the reverse system, customers enter the network at the points where they leave in the forward system. Once in the network, they travel backwards along the same path as they would take through the forward network, and leave the network via the point at which they enter in the forward system ([10, p. 69]). These ideas can be summarized by saying that a forward-time arrival is a reverse-time service. The arrival process is poisson in both forward and reverse networks.

Using the idea of customers traveling backwards through the network, [10] argues that at each queue in the network, the total arrival rate is the same in the forward and reverse systems, (i.e., $\gamma_{iR} = \gamma_i$,) and that the same is true of the virtual service rate (i.e., $\mu_{iR} = \mu_i$.) The external arrival rate at queue i) in the reverse system (λ_{iR} is the rate at which customers leave the network at queue i in the forward system. Hence, $\lambda_{iR} = \gamma_i p_{i0}$. Finally, the rate at which customers travel along all paths of the network is the same in the forward and reverse systems, so $p_{ijR}\gamma_{iR} = p_{ji}\gamma_j$.

These results for generating the reverse-time model of a Jackson network can be summarized by the following equations:

$$\gamma_{iR} = \gamma_i \tag{37a}$$

$$\mu_{iR} = \mu_i \tag{37b}$$

$$\lambda_{iR} = \gamma p_{i0} \tag{37c}$$

$$p_{ijR} = \frac{\gamma_j p_{ji}}{\gamma_i}. \tag{37d}$$

A formal justification for this method of constructing the reverse network is given in [10].

B. Single $M/M/1$ Queue

Imagine an $M/M/1$ queue with arrival rate λ, service rate μ, and buffer size N. Let us say that the queue has been running for a long time, and that its buffer has just overflowed. We can look at the period leading up to a buffer overflow by running the queue's reverse-time model backwards in time, starting with its buffer full. The forward and reverse systems are shown in Fig. 4.

If the reverse-system is run backwards in time, the number of customers in the queue will on average decay with rate $(\mu - \lambda)$. During this period, corresponding to that leading up to an overflow in the forward system, the buffer is never empty, and hence the average rate of reverse-time services is μ, and the average rate of reverse-time arrivals is λ. Reverse-time arrivals are forward time services and reverse-time services are forward-time arrivals. Hence, during the period leading up to an overflow, the average behavior of the forward system is as if had an arrival rate μ and service rate λ. This is the same as the average behavior of the system $\bar{S}$, which has arrival rate μ and service rate λ, as shown in Fig. 2.

C. Jackson Networks

A similar statement to that made in the previous section for $M/M/1$ queues can be made about the relationship between S and $\bar{S}$ for Jackson networks. In this case, if an overflow has just occurred, then there are N customers in the network. If we look at the behavior of the dominating queue, we expect that it will contain the majority of these customers. Without loss of generality, we will assume that queue 1 is this dominating queue. It will be shown that the average behavior of the original network in the period ending in an overflow, and beginning in the last time before this that the dominating queue was empty, is the same as the average behavior of the simulation system $\bar{S}$.

We now consider the time back until this queue was last empty, using the reverse-time model. If we run the reverse-time model of the network backwards in time from the full-buffer state to the empty-buffer state, then the rate of customers leaving the dominating queue is actually μ_1, this being the virtual service rate, since the queue is never empty. This is the same as the behavior of the $M/M/1$ queue, and corresponds to an arrival rate, over the interval under consideration, equal to μ_1 in the forward system. If there is no feedback around the dominating queue, (i.e., $r_1 = 1$) then the rate of customers entering this queue in the reverse

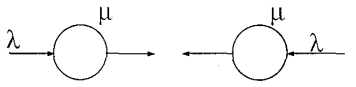


Fig. 4. An $M/M/1$ queue and its reverse-time model.

system is γ_1, which is also the same as the behavior for the isolated $M/M/1$ queue. If there is feedback in the network, then we must correct this figure: the arrivals are made up of two streams, one of customers that have never passed through queue 1 before, and a second stream comprising a fixed proportion of those customers that have passed through queue 1 before. The component of the arrival rate attributable to the second stream must be recomputed to reflect the fact that over the time interval in question, the average exit rate, at μ_1, exceeds the arrival rate to the queue. Let π be the probability that a customer exiting queue 1 in the reverse system is fed back to the input, so that $r_1 = (1 - \pi)^{-1}$. Then the arrival rates of the two streams under normal circumstances are $(1 - \pi)\gamma_1$ and $\pi\gamma_1$, respectively. However, with the exit rate raised to μ_1, the arrival rates become $(1 - \pi)\gamma_1$, and $\pi\mu_1$, i.e., the effective reverse-time arrival rate, or forward-time service rate, is

$$\mu_{1\,\text{eff}} = (1 - \pi)\gamma_1 + \pi\mu_1 \tag{38}$$

$$= \frac{\gamma_1}{r_1} + \left(1 - \frac{1}{r_1}\right)\mu_1 \tag{39}$$

$$= \gamma_1 + \frac{(r_1 - 1)(\mu_1 - \gamma_1)}{r_1}. \tag{40}$$

In summary, the dominating queue, over the interval of interest, behaves on average as if its arrival rate were $\gamma_{1\,\text{eff}} = \mu_1$, and as if its service rate were $\mu_{1\,\text{eff}} = \gamma_1 + (r_1 - 1)(\mu_1 - \gamma_1)/r_1$, which are the same average rates as occur in the optimal simulation system defined above.

Since the average behavior of other queues does not result in a build up, the average virtual service rates at other queues remain unchanged in the period leading up to an overflow.

During this period leading up to an overflow, the rates at all external outputs from the system are on average the same as in the average behavior of the system over all time. Hence, we have (27d), if we equate primed terms in (27d) with the behavior of the system leading up to an overflow.

Also, in this period leading up to an overflow, the total arrival rate at queues other than queue 1 is increased in proportion to its contribution to γ_1 and to the relative increase in γ_1, i.e., $\gamma_{1\,\text{eff}} - \gamma_1/\gamma_1$. Finally, in this period, the p_{ij} on average change their values such that the ratio of the rates in two paths ending at the same queue is that same as it is in the infinite-time average case, as in (27e).

Hence, it can be seen that the average behavior of the network, in the period between an overflow occurring and the immediately previous time that the dominating queue was empty, is the same as the average behavior of the simulation system.

V. Conclusion

This paper has extended the previously known theory for generating optimal (in the sense of variance) importance sampling simulation systems in two areas. First, a simple analytic method has been derived for the construction of such systems, removing the need to perform numerical minimizations. This derivation did not require the assumption of exponential service rates on queues other than that dominating the overflow statistics. In some cases, it is also possible to remove the assumption that arrival streams are poisson. Second, it has been shown that the average behavior original system in the time leading up to an overflow is the same as the average behavior of this asymptotically optimal simulation system.

This work could be extended further by finding the optimal simulation system for networks where the arrival streams are not poisson, and to the case where the service times of the dominating queue are not exponentially distributed.

Appendix A
Derivation of Cost Function

We will now present the heuristic derivation [6] for the optimal simulation system of $GI/GI/1$ queue. This derivation is based on a heuristic due to Borovkov, Ruget, and others. The basic idea behind this derivation is that for fast simulation we should use a system whose average path is the same as the average path that is followed in the original system *immediately leading up to an escape.*

We consider a $GI/GI/1$ queue with interarrival time distribution A, and virtual service time distribution B. Let $1/\lambda$ be the mean of A, and $h_A(\cdot)$ its Cramér transform. Similarly, let $1/\mu$ be the mean of B, and $h_B(\cdot)$ its Cramér transform. We assume that $\lambda < \mu$ for stability. As usual, we wish to find by simulation the probability α that a cycle ends in an overflow. Let X_i^A denote the ith interarrival time and X_i^B the ith virtual service time.

We define a cycle as a piece of trajectory starting with the buffer empty and terminating the first time that the buffer is empty again, or an overflow occurs. We will consider only those cycles that terminate in an overflow. Let the average interarrival time in these cycles of the second kind be $1/\lambda'$, and the average virtual service time be $1/\mu'$. Then by Cramér's theorem over a time T

$$P\{X_1^A + \cdots + X_{\lambda' T}^A \simeq T\}$$

$$= P\left\{\frac{X_1^A + \cdots + X_{\lambda' T}^A}{\lambda' T} \simeq \frac{1}{\lambda'}\right\}$$

$$\simeq \exp\left(-\lambda' T h_A\left(\frac{1}{\lambda'}\right)\right) \quad \text{(UTLE)} \tag{41}$$

$$P\{X_1^B + \cdots + X_{\mu' T}^B \simeq T\}$$

$$= P\left\{\frac{X_1^B + \cdots + X_{\mu' T}^B}{\mu' T} \simeq \frac{1}{\mu'}\right\}$$

$$\simeq \exp\left(-\mu' T h_B\left(\frac{1}{\mu'}\right)\right) \quad \text{(UTLE)} \tag{42}$$

where "UTLE" is an abbreviation for "up to logarithmic equivalence."

Since we examine only the one cycle immediately before an overflow occurs, all virtual services are actual services, and $T = N/\lambda' - \mu'$. Because the arrivals and services are independent, it can be claimed that [6] (using the notation of [6])

$$\alpha \simeq \sum_T \sum_{\substack{\lambda' > \mu' > 0 \\ N = T(\lambda' - \mu')}} \exp\left\{-T\left(\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right)\right\}$$

$$= \sum_{\lambda' > \mu' > 0} \exp\left\{-\frac{N}{\lambda' - \mu'}\left(\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right)\right\}. \quad (43)$$

Hence, for large N

$$\alpha \sim \exp\left\{-N \inf_{\lambda' > \mu' > 0} \frac{1}{\lambda' - \mu'}\left(\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right)\right\} \quad (\text{UTLE}). \quad (44)$$

By performing the minimization, it can be shown that the minimum satisfies [6]

$$h_A\left(\frac{1}{\lambda'}\right) + h_B\left(\frac{1}{\mu'}\right) = \left(\frac{1}{\lambda'} - \frac{1}{\mu'}\right) h'_A\left(\frac{1}{\lambda'}\right) \quad (45)$$

$$= \left(\frac{1}{\mu'} - \frac{1}{\lambda'}\right) h'_B\left(\frac{1}{\mu'}\right). \quad (46)$$

The values of λ' and μ' satisfying these equations are the average arrival and service rates for the optimal simulation system.

Appendix B
Proof of Optimality

In this section, it is shown that the set of equations given in Section III defines a global minimum of the exponent of (21). First, it will be shown that these equations correspond to a turning point, and then that this point is in fact the required minimum. As in the previous text, we will use the symbol S to represent the original system and $\bar{S}$ to represent the importance-sampling system generated by large deviations.

B.1. Evaluation of Derivatives of Lagrangian

Let

$$H = R\left[\sum_{i=1}^{d} \lambda'_i h_{\lambda_i}\left(\frac{1}{\lambda'_i}\right) + \sum_{i=1}^{d} \mu'_i h_{\mu_i}\left(\frac{1}{\mu'_i}\right) + \sum_{i=1}^{d} \min\left(\gamma'_i, \mu'_i\right) K_i\right] \quad (47)$$

where R and K_i are as defined in the main text. Then the extrema of H subject to the equality constraints (24d) through (24f) are found by setting the partial derivatives of the lagrangian to zero; the relevant equations are

$$\frac{\partial \mathcal{L}}{\partial \lambda'_i} = R \log \frac{\lambda'_i}{\lambda_i} + g + b_i = 0 \quad (48)$$

$$\frac{\partial \mathcal{L}}{\partial \mu'_i} = \begin{cases} RH + R\left[\log \dfrac{\mu'_i}{\mu_i} + \displaystyle\sum_{j=0}^{d} p'_{ij} \log \frac{p'_{ij}}{p_{ij}}\right] \\ \quad + g + \displaystyle\sum_{j=1}^{d} b_j p'_{ij} \qquad \gamma'_i > \mu'_i \\ R \log \dfrac{\mu'_i}{\mu_i} + g \qquad \text{otherwise} \end{cases} = 0 \quad (49)$$

$$\frac{\partial \mathcal{L}}{\partial p'_{i0}} = R \min\left(\gamma'_i, \mu'_i\right)\left[1 + \log \frac{p'_{i0}}{p_{i0}}\right] + c_i = 0 \quad (50)$$

$$\frac{\partial \mathcal{L}}{\partial p'_{ij}} = R \min\left(\gamma'_i, \mu'_i\right)\left[1 + \log \frac{p'_{ij}}{p_{ij}}\right] + b_j \min\left(\gamma'_i, \mu'_i\right) + c_i = 0 \qquad \text{for } j \geq 1 \quad (51)$$

$$\frac{\partial \mathcal{L}}{\partial \gamma'_i} = \begin{cases} -RH - b_i & \gamma'_i > \mu'_i \\ R \displaystyle\sum_{j=0}^{d} p'_{ij} \log \frac{p'_{ij}}{p_{ij}} + \sum_{j=1}^{d} b_j p'_{ij} - b_i & \text{otherwise} \end{cases} = 0 \quad (52)$$

$$\frac{\partial \mathcal{L}}{\partial g} = \sum_{i=1}^{d} \left(\lambda_i + \mu'_i\right) - 1 = 0 \quad (53)$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \lambda_i + \sum_{j=1}^{d} \min\left(\gamma'_j, \mu'_j\right) p'_{ji} - \gamma'_i = 0 \quad (54)$$

$$\frac{\partial \mathcal{L}}{\partial c_i} = \sum_{j=0}^{d} p'_{ij} - 1 = 0. \quad (55)$$

It can be shown by direct substitution of (27a) through (27e) and (28a) through (28c) that these equations are satisfied by the solution outlined in Section III. Therefore, it is clear that these equations define at least a turning point of H, under the equality constraints in (24d) through (24f). It will be shown later that this solution also satisfies the inequality constraints (24a) through (24c). In the following sections it will be shown that this solution is in fact a global minimum.

B.2. Evaluation of Optimal Simulation System

Given a system $\bar{S}(\lambda'_i, \mu'_i, \gamma'_i, p'_{ij})$ whose parameters define a turning point of H, we assume without loss of generality that I queues are unstable in $\bar{S}$, and that (after renumbering if necessary) the queues are numbered such that queues 1 through I are unstable and all others are stable. For the moment, we do not suppose that under this numbering, queue 1 is the most heavily loaded.

Let r_{ij} be the expected number of times that a customer arriving at queue i will pass through queue j before leaving the network, and let $\boldsymbol{r} = (r_{ij})$ and $\boldsymbol{p} = (p_{ij})$ for $i, j \in [1, d]$, (i.e., $\boldsymbol{p}$ and $\boldsymbol{r}$ are square matrices of dimension d).

Just as in the main text, where r_i was the value of γ_1 when $\lambda_i = 1$ and $\lambda_j = 0$ for $j \neq 0$, here we have that r_{ij} is

the value of γ_j when $\lambda_i = 1$ and $\lambda_k = 0$ for $k \neq i$. Hence

$$\gamma_j = \sum_{i=1}^{d} r_{ij}\lambda_i \tag{56}$$

and

$$\boldsymbol{r} = \boldsymbol{I} = \boldsymbol{pr}. \tag{57}$$

Therefore

$$(\boldsymbol{I} - \boldsymbol{p})\boldsymbol{r} = \boldsymbol{I}. \tag{58}$$

As has been argued previously for tandem networks of queues [7], physical constraints require that $g = 0$. The reason for this is that g is the rate of change of $\mathscr{L}$ with the sum of the arrival and service rates, i.e., the scaling of time, and we do not expect the probability that a cycle exits rather than returns to zero to depend on the scaling of time. Therefore, we require $g = 0$. Then (49) yields $\mu_i' = \mu_i$ for $i > I$.

Using (48) to substitute for b_i in (51), and using also (50), we have for $i, j \in [1, d]$

$$\frac{p_{ij}'}{p_{ij}} = \frac{\lambda_j'}{\lambda_j}\frac{p_{i0}'}{p_{i0}}. \tag{59}$$

Now consider the first equation of (49). Recognize from (52) that $RH = -b_i = R \log \lambda_i'/\lambda_i$ for $i \in [1, I]$. Substitute also for p_{ij}'/p_{ij} using (59). There results the first equation of (60) below, after simplifying. In a similar way, using the second equation of (52), and substituting for b_i and p_{ij}'/p_{ij}, the second equation of (60) results: (49) and (52), we obtain

$$\frac{p_{i0}'}{p_{i0}} = \begin{cases} \dfrac{\lambda_i}{\lambda_i'}\dfrac{\mu_i}{\mu_i'} & i \leq I \\ \dfrac{\lambda_i}{\lambda_i'} & i > I. \end{cases} \tag{60}$$

Hence from (59), for $j \geq 1$

$$\frac{p_{ij}'}{p_{ij}} = \begin{cases} \dfrac{\lambda_j'}{\lambda_j}\dfrac{\lambda_i}{\lambda_i'}\dfrac{\mu_i}{\mu_i'} & i \leq I \\ \dfrac{\lambda_j'}{\lambda_j}\dfrac{\lambda_i}{\lambda_i'} & i > I. \end{cases} \tag{61}$$

(We observe that these equations are consistent with the solution of the minimization problem set out in the main text, where $I = 1$.) Replacing p_{ij}' in (55) with the expansion available in (60) and (61) gives:

$$(\boldsymbol{I} - \boldsymbol{p})\begin{pmatrix} \dfrac{\lambda_1'}{\lambda_1} \\ \dfrac{\lambda_2'}{\lambda_2} \\ \vdots \\ \dfrac{\lambda_d'}{\lambda_d} \end{pmatrix} = \begin{pmatrix} p_{10} \\ p_{20} \\ \vdots \\ p_{d0} \end{pmatrix} + \begin{pmatrix} \dfrac{\lambda_1'}{\lambda_1}\left(1 - \dfrac{\mu_i'}{\mu_1}\right) \\ \vdots \\ \dfrac{\lambda_I'}{\lambda_I}\left(1 - \dfrac{\mu_I'}{\mu_I}\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{62}$$

Also, from (20), we have

$$\begin{pmatrix} P_{10} \\ p_{20} \\ \vdots \\ p_{d0} \end{pmatrix} = (\boldsymbol{I} - \boldsymbol{p})\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{63}$$

Eliminating the p_{i0} terms from (62), using (63), and substituting $\boldsymbol{r}$ for $(\boldsymbol{I} - \boldsymbol{p})^{-1}$, we obtain

$$\frac{\lambda_i'}{\lambda_i} = 1 + \sum_{j=1}^{I} r_{ij}\frac{\lambda_j'}{\lambda_j}\left(1 - \frac{\mu_j'}{\mu_j}\right). \tag{64}$$

From (49), for queues with index $i \leq I$, we obtain

$$0 = RH + R\left[\log\frac{\mu_i'}{\mu_i} + \sum_{j=0}^{d} p_{ij}' \log\frac{p_{ij}'}{p_{ij}}\right] + g + \sum_{j=1}^{d} b_j p_{ij}' \tag{65}$$

$$= R\left[\log\frac{\lambda_i'}{\lambda_i} + \log\frac{\mu_i'}{\mu_i} + p_{i0}' \log\frac{p_{i0}'}{p_{i0}} + \sum_{j=1}^{d}\left(p_{ij}' \log\frac{p_{ij}'}{p_{ij}}\frac{\lambda_j}{\lambda_j'}\right)\right] \tag{66}$$

$$= R\left[\sum_{j=1}^{d}\left(p_{ij}' \log\frac{p_{ij}'}{p_{ij}}\frac{\lambda_j}{\lambda_j'}\right)\right] \quad \text{from (60)} \tag{67}$$

$$= R(1 - p_{i0}')\log\frac{p_{i0}'}{p_{i0}} \quad \text{from (55) and (59)}, \tag{68}$$

i.e., since R cannot be zero, for $i \in [1, I]$, $p_{i0}' = p_{i0}$, or $p_{i0}' = 1$. Clearly, it is the first solution in which we are interested, since $p_{i0}' = 1$ implies $p_{ij}' = 0$ for all $j > 0$, and hence from (59), that $\lambda_j = 0$ for all queues to which customers can be routed in one step from queue i. Therefore, from (60), we have

$$\frac{\mu_i'}{\mu_i} = \frac{\lambda_i}{\lambda_i'} = \text{constant} \qquad \forall i \leq I, \tag{69}$$

i.e., for all queues that are unstable in $\bar{S}$, the ratio μ_i'/μ_i takes the same value.

Now, in order for (53) to hold, we must have

$$\begin{aligned} 0 &= \sum_{i=1}^{d}(\lambda_i + \mu_i') - 1 \\ &= \sum_{i=1}^{d}\lambda_i + \sum_{i=I+1}^{d}\mu_i + \sum_{j=1}^{d}\sum_{i=1}^{I} r_{ji}\lambda_j\frac{\lambda_i'}{\lambda_i}\left(1 - \frac{\mu_i'}{\mu_i}\right) \\ &\quad + \sum_{i=1}^{I}\mu_i' - 1 \\ &= \sum_{i=1}^{I}\gamma_i\frac{\lambda_i'}{\lambda_i}\left(1 - \frac{\mu_i'}{\mu_i}\right) - \sum_{i=1}^{I}\mu_i\left(1 - \frac{\mu_i'}{\mu_i}\right). \end{aligned} \tag{70}$$

With a small amount of manipulation, in particular using the

fact that $\lambda_i'/\lambda_i = \mu_i/\mu_i'$ is constant for $i \le I$, it can be shown that

$$\lambda_i' = \lambda_i\left(\frac{\mu_1 + \cdots + \mu_I}{\gamma_1 + \cdots + \gamma_I}\right) \qquad \text{for } i \le I. \tag{71}$$

Therefore, assuming that a solution exists for which there are I unstable queues in $\bar{S}$, as postulated above, and substituting for b_i from (48) in (52) for $i \le I$, we have

$$H = \log\left(\frac{\mu_1 + \cdots + \mu_I}{\gamma_1 + \cdots + \gamma_I}\right). \tag{72}$$

The postulate that there are a particular I unstable queues in $\bar{S}$ corresponding to a turning point of H has led to the conclusion that only one set of values $(\lambda_i', \mu_i', \gamma_i', p_{ij}')$ apparently give a turning point; note, however, that for these values to actually give a turning point, there would have to be satisfied at the solution point the conditions $\gamma_i' > \mu_i'$ for $1 \le i \le I$, as well as the other inequality conditions in (24). If these conditions are not satisfied, the postulate that a particular I queues are unstable is inconsistent with there being an associated turning point of H.

Now consider all possible values of I, and all possible selections of I queues. With each such selection, there is a potential turning point for H (which will actually be a turning point only if the postulated instability condition is actually fulfilled at the solution point). The set of associated values of H has a minimum element, obtained by choosing only the queue with the highest load in S to be unstable in $\bar{S}$, since if $\gamma_1/\mu_1 > \gamma_i/\mu_i$ for all $i \ne 1$

$$\log\frac{\mu_1}{\gamma_1} < \log\left(\frac{\mu_{i_1} + \cdots + \mu_{i_I}}{\gamma_{i_1} + \cdots + \gamma_{i_I}}\right) \tag{73}$$

unless $i_1 = 1$ and $i_2 \cdots i_I$ are empty.

The solution point obtained above for the assumption that queue 1, and no other queue, is unstable in $\bar{S}$ turns out to always lead to queue 1, and no other queue, being unstable; this is established in the text. Further, the other inequality constraints in (24) hold; this is established in the next section. Hence, the minimum of the set of possible values of H is actually attained.

B.3. Satisfaction of Inequality Constraints

The use of the method of Lagrange multipliers ensures that the equality constraints of (24) are satisfied. We have already shown in the main text that the third of the inequality constraints, requiring that at least one queue in the simulation system becomes unstable, is satisfied by the solution presented here. (In fact, we showed that just one queue is unstable.) It remains to be shown that $\lambda_i', \mu_i', \gamma_i' \ge 0$ and $0 < p_{ij}' \le 1$.

First, it is clear from (27a) that $\gamma_i' > 0\ \forall i$, since we know that all queues in S are stable (i.e., $\mu_i > \gamma_i$.) Hence, (27b) shows that $\lambda_i' > 0\ \forall i$. From (27c), because $r_1 \ge 1$, it is clear that we must have $\mu_i' > 0\ \forall i$.

Given the above, (27d) and (27e) imply $p_{ij}' \ge 0$. The requirement $p_{ij}' \le 1$ is enforced by the equality constraint $\sum_j p_{ij}' = 1$. Hence, all the inequality constraints are satisfied.

B.4. Generalization of Distribution Functions

As we have argued previously [7], there is no requirement that the distribution of service times for queues other queue 1 be exponential. By definition, the subscript μ of $h_\mu(\cdot)$ is the expected number of services per unit time. Hence, the mean time between services is $1/\mu$. Therefore, for buffers whose service rates are the same in both the original and optimal-simulation systems, (i.e., $\mu_j = \mu_j'$) we have

$$h_{\mu_j}\left(\frac{1}{\mu_j'}\right) = h_{\mu_j}\left(\frac{1}{\mu_j}\right) = 0 \tag{74}$$

since the Cramér transform has value zero at the mean of its associated distribution function.

It is guaranteed that $h(\cdot)$ is zero at exactly one point, that point being the mean of the associated distribution. At that point, its derivative is also zero [3]. Therefore, for buffers other than that which dominates the overflow statistics, (49) can be written

$$\begin{aligned} 0 &= \frac{\partial}{\partial\mu_j'}\frac{1}{\gamma_1' - \mu_1'}h_{\mu_j}\left(\frac{1}{\mu_j'}\right) \\ &= \frac{1}{\gamma_1' - \mu_1'}\frac{\partial}{\partial\mu_j'}h_{\mu_j}\left(\frac{1}{\mu_j'}\right) \end{aligned} \tag{75}$$

where $j \ne i$. This implies that the derivatives of $h_{\mu_j}(1/\mu_j')$ with respect to μ_j' is zero at $\mu_j' = \mu_j$. Hence, it is clear that for queues other than that which dominates the overflow statistics $\mu_j' = \mu_j$ *regardless of the distribution associated with these service rates.*

An identical argument shows that any queue i for which $r_i = 0$, $\lambda_i' = \lambda_i$, and that this does not depend on the distribution of these external arrival rates.

Hence, we see that for sufficiently large N, the optimal simulation system depends only on the statistics of the service rate of one queue (that of the least serviced buffer) and the arrival process, assuming that no two service rates are actually equal, and does not depend in any way on the statistics of the service rates of buffers other than the one dominating the overflow statistics.

References

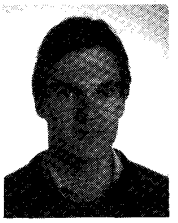
[1] M. I. Friedlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*. New York: Springer-Verlag, 1984.

[2] J. Zabczyk, "Exit problem and control theory," *Syst. Contr. Lett.*, vol. 6, pp. 165–172, Aug. 1985.

[3] M. Cottrell, J. C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 907–918, Sept. 1983.

[4] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*. Baltimore, MD: Johns Hopkins University Press, 1981.

[5] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[6] S. Parekh and J. Walrand, "A quick simulation of excessive backlogs in networks of queues," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 54-66, January 1989.
[7] M. R. Frater and B. D. O. Anderson, "Fast estimation of the statistics of excessive backlogs in tandem networks of queues," *Australian Telecommunication Research*, vol. 23, pp. 49-55, May 1989.
[8] M. R. Frater, R. A. Kennedy, and B. D. O. Anderson, "Reverse-time modeling, optimal control and large deviations," *Syst. Contr. Lett.*, vol. 12, pp. 351-356, May 1989.
[9] M. R. Frater, "Fast estimation of the statistics of rare events in data communications systems," Ph.D. dissertation, Australian National University, Nov. 1990.
[10] F. P. Kelly, *Reversibility and Stochastic Processes*. New York: Wiley, 1979.
[11] I. Mitrani, *Modelling of Computer and Communication Systems*. New York: Cambridge University Press, 1987.
[12] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.



Michael R. Frater was born in Sydney, Australia in 1965. He received the B.Sc. degree in mathematics and physics in 1986 and the B.E. (Hons.) degree in electrical engineering in 1988 from the University of Sydney. He received the Ph.D. degree in the Department of Systems Engineering at the Australian National University, Canberra, in 1991.

He is currently a lecturer in electrical engineering at the University College, Australian Defence Force Academy. His research interests include teletraffic and queueing theory, stochastic processes (especially problems relating to large deviations), signal processing and control.

Tava M. Lennon, photograph and biography not available at time of publication.

Brian D. O. Anderson (S'62-M'66-SM'74-F'75), for a photograph and biography, see p. 473 of the April 1991 issue of this TRANSACTIONS.